Embedding Methods for NLP

Part 2: Embeddings for Multi-relational Data

Antoine Bordes
Facebook AI Research

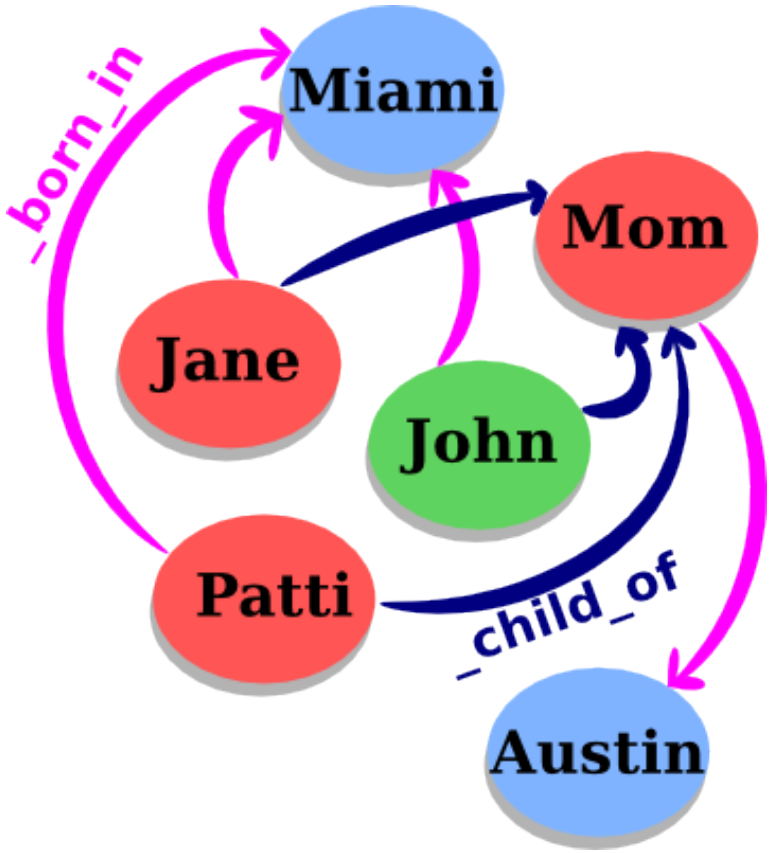
NYU Deep Learning Lectures – April 13, 2015

Menu – Part 2

Menu – Part 2

Multi-relational data

- Data is structured as a graph
- Each node = an entity
- Each edge = a relation/fact
- A relation = (*sub*, *rel*, *obj*):
  - *sub* =*subject*,
  - *rel* = *relation type*,
  - *obj* = *object*.
- Nodes w/o features.

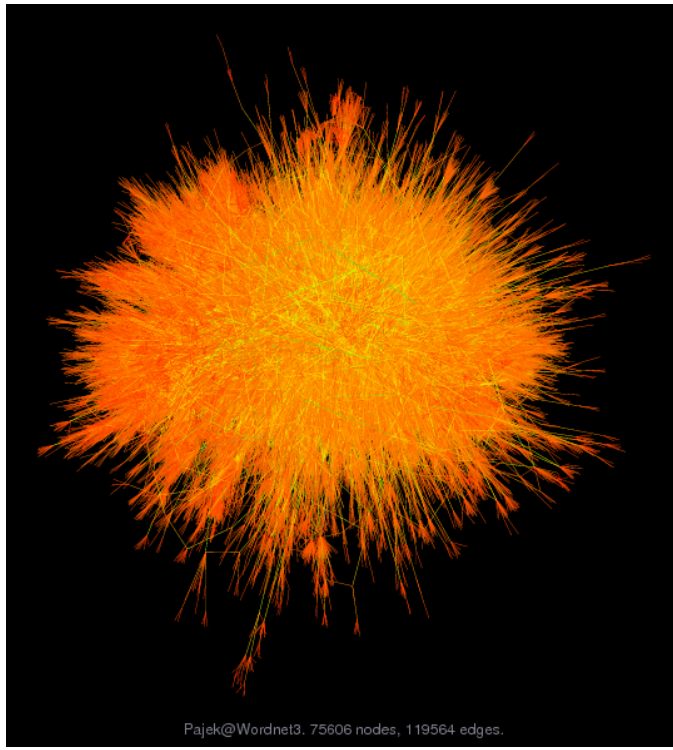


In this talk, we focus on Knowledge Bases (KBs).

Example of KB: WordNet

- WordNet: dictionary where each entity is a sense (synset).

- Popular in NLP.
- Statistics:
  - 117k entities;
  - 20 relation types;
  - 500k facts.
- Examples:
  (car_NN_1, _has_part, _wheel_NN_1)
  (score_NN_1, _is_a, _rating_NN_1)
  (score_NN_2, _is_a, _sheet_music_NN_1)



Example of KB: Freebase

- Freebase: huge collaborative (hence noisy) KB.
- Part of the Google Knowledge Graph.
- Statistics:
  - 80M of entities;
  - 20k relation types;
  - 1.2B facts.

- Examples:
  (Barack Obama, _place_of_birth, Hawai)
  (Albert Einstein, _follows_diet, Veganism)
  (San Francisco, _contains, Telegraph Hill)

Modeling Knowledge Bases

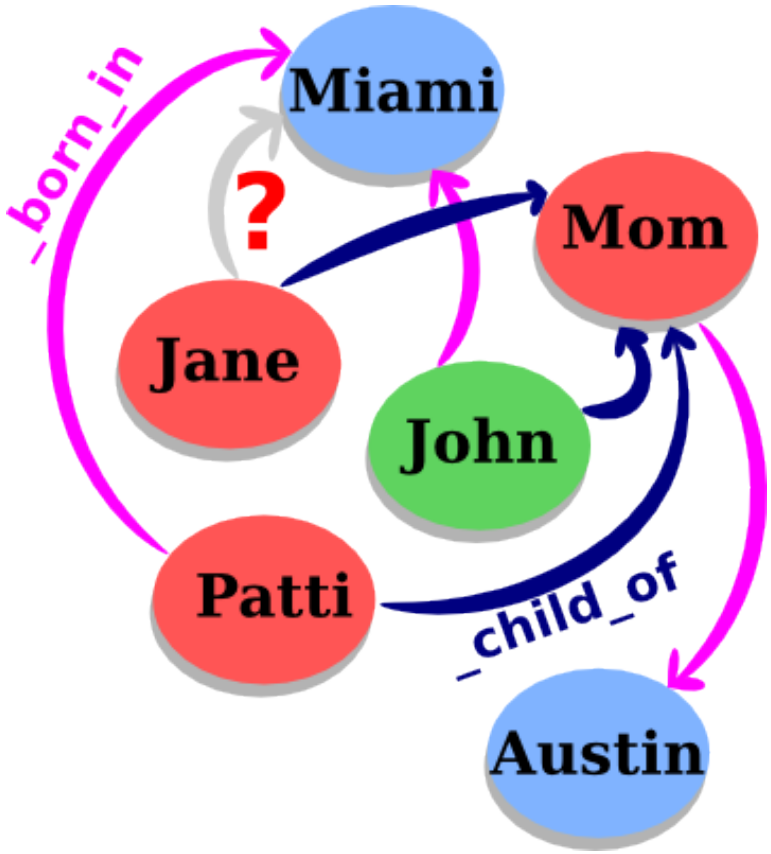
- **Why KBs?**
  - KBs: Semantic search, connect people and things
  - KBs $\leftarrow$ Text: Information extraction
  - KBs $\rightarrow$ Text: Text interpretation, summary, Q&A
- **Main issue:** KBs are hard to manipulate
  - Large dimensions: $10^5/10^8$ entities, $10^4/10^6$ rel. types
  - Sparse: few valid links
  - Noisy/incomplete: missing/wrong relations/entities
- **How?**
  1. Encode KBs into low-dimensional vector spaces
  2. Use these representations:
     - to complete/visualize KBs
     - as KB data in text applications

Menu – Part 2

Link Prediction

Add new facts without requiring extra knowledge

From known information, assess the validity of an unknown fact

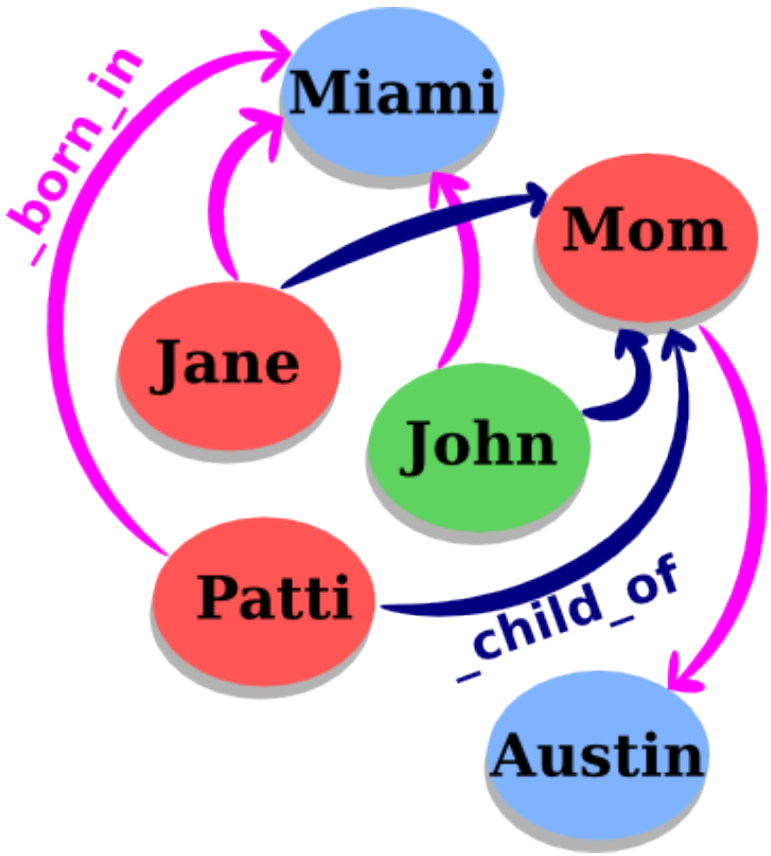


Link Prediction

Add new facts without requiring extra knowledge

From known information, assess the validity of an unknown fact

→ *collective classification*
→ *reasoning in embedding spaces*



Statistical Relational Learning

- **Framework:**
  - n_s subjects $\{sub_i\}_{i\in[1;n_s]}$
  - n_r relation types $\{rel_k\}_{k\in[1;n_r]}$
  - n_o objects $\{obj_j\}_{j\in[1;n_o]}$

  $\rightarrow$ For us, $n_s = n_o = n_e$ and $\forall i \in [1; n_e]\,, sub_i = obj_i$.

  - A fact exists for (sub_i, rel_k, obj_j) if $rel_k(sub_i, obj_j) = 1$
- **Goal:** We want to model, from data,

$$\mathbb{P}[rel_k(sub_i, obj_j) = 1]$$

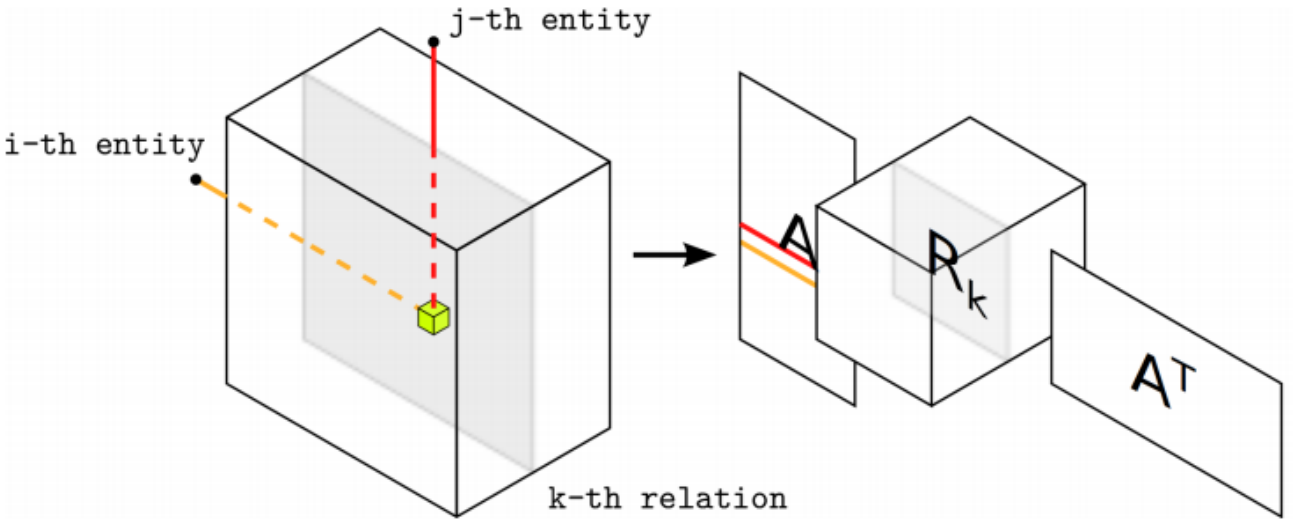
(eq. to approximate the binary tensor $\mathbf{X} \in \{0,1\}^{n_s \times n_o \times n_r}$)

Previous Work

- Tensor factorization (Harshman et al., '94)
- Probabilistic Relational Learning (Friedman et al., '99)
- Relational Markov Networks (Taskar et al., '02)
- Markov-logic Networks (Kok et al., '07)
- Extension of SBMs (Kemp et al., '06) (Sutskever et al., '10)
- Spectral clustering (undirected graphs) (Dong et al., '12)
- Ranking of random walks (Lao et al., '11)
- Collective matrix factorization (Nickel et al., '11)
- Embedding models (Bordes et al., '11, '13) (Jenatton et al., '12) (Socher et al., '13) (Wang et al., '14) (García-Durán et al., '14)

Collective Matrix Factorization (Nickel et al., '11)

- RESCAL: $\forall k \in [1; n_r], \mathbf{R}_k \in \mathbb{R}^{d \times d}$ and $\mathbf{A} \in \mathbb{R}^{n_e \times d}$ (close from DEDICOM (Harshman, '78)).



- **A** & **R** learned by reconstruction (alternating least-squares):

$$\min_{\mathbf{A},\mathbf{R}} \frac{1}{2} \left(\sum_k ||\mathbf{X}_k - \mathbf{A}\mathbf{R}_k\mathbf{A}^\top||_F^2 \right) + \lambda_A ||\mathbf{A}||_F^2 + \lambda_R \sum_k ||\mathbf{R}_k||_F^2$$

Scalability

Method	Nb of parameters	on Freebase15k
RESCAL	$O(n_e d + n_r d^2)$	88M ($d = 250$)

Freebase15k: $n_e = 15k$, $n_r = 1.3k$.

- RESCAL involves many parameters.
- Bad scalability w.r.t. n_r.
- Reconstruction criterion does not fit well for binary data..

Embedding Models

Two main ideas:

1. Models based on low-dimensional continuous vector embeddings for entities and relation types, directly trained to define a similarity criterion.

2. Stochastic training based on ranking loss with sub-sampling of unknown relations.

Embedding Models for KBs

- Subjects and objects are represented by vectors in $\mathbb{R}^d$.
  - $\{sub_i\}_{i\in[1;n_s]} \quad \rightarrow \quad [\mathbf{s}^1, \ldots, \mathbf{s}^{n_s}] \in \mathbb{R}^{d\times n_s}$
  - $\{obj_j\}_{j\in[1;n_o]} \quad \rightarrow \quad [\mathbf{o}^1, \ldots, \mathbf{o}^{n_o}] \in \mathbb{R}^{d\times n_o}$

  For us, $n_s = n_o = n_e$ and $\forall i \in [1; n_e], \mathbf{s}_i = \mathbf{o}_i$.

- **Rel. types = similarity operators between subj/obj.**
  - $\{rel_k\}_{k\in[1;n_r]} \quad \rightarrow \quad$ operators $\{\mathbf{R}_k\}_{k\in[1;n_r]}$

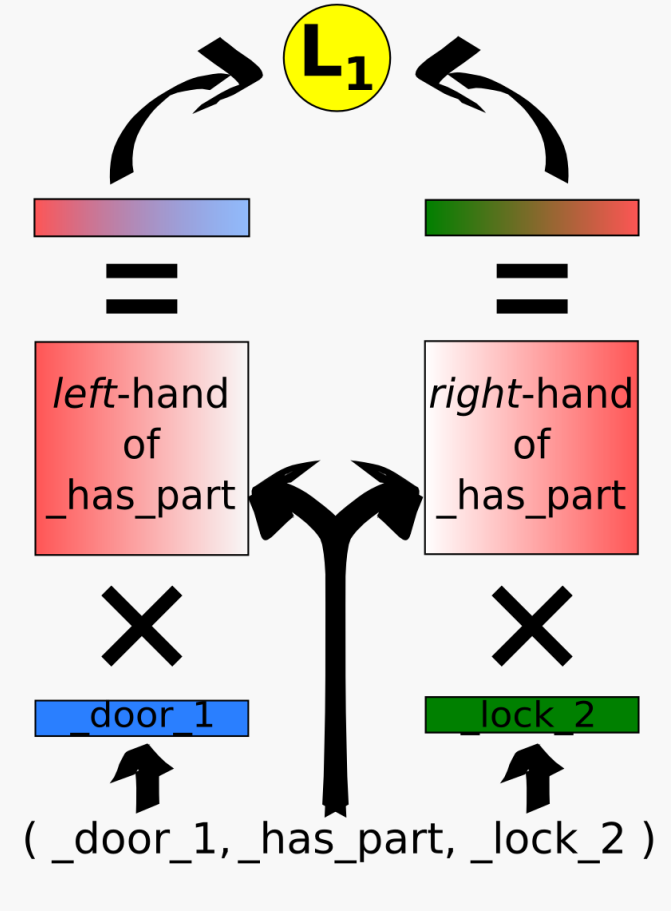
- Learning similarities depending on $rel \rightarrow d(sub, rel, obj)$, parameterized by $\mathbf{s}$, $\mathbf{R}$ and $\mathbf{o}$.

Structured Embeddings (Bordes et al., '11)

Intuition: *sub* and *obj* are projected using *rel* in a space where they are similar

$d(sub, rel, obj) = -||\mathbf{R}^{left}\mathbf{s}^\top - \mathbf{R}^{right}\mathbf{o}^\top||_1$

- Entities: $\mathbf{s}$ and $\mathbf{o} \in \mathbb{R}^d$
- Projection: $\mathbf{R}^{left}$ and $\mathbf{R}^{right} \in \mathbb{R}^{d \times d}$
  $\mathbf{R}^{left} \neq \mathbf{R}^{right}$ because of asymmetry
- Similarity: L1 distance



Stochastic Training

- Learning by stochastic gradient descent: one training fact after the other
- For each relation from the training set:
  1. sub-sample unobserved facts (false?)
  2. check if the similarity of the true fact is lower
  3. **if not**, update parameters of the considered facts
- Stopping criterion: performance on a validation set

Scalability

Method	Nb of parameters	on Freebase15k
RESCAL	$O(n_e d + n_r d^2)$	88M ($d = 250$)
SE	$O(n_e d + 2n_r d^2)$	8M ($d = 50$)

Freebase15k: $n_e = 15k$, $n_r = 1.3k$.

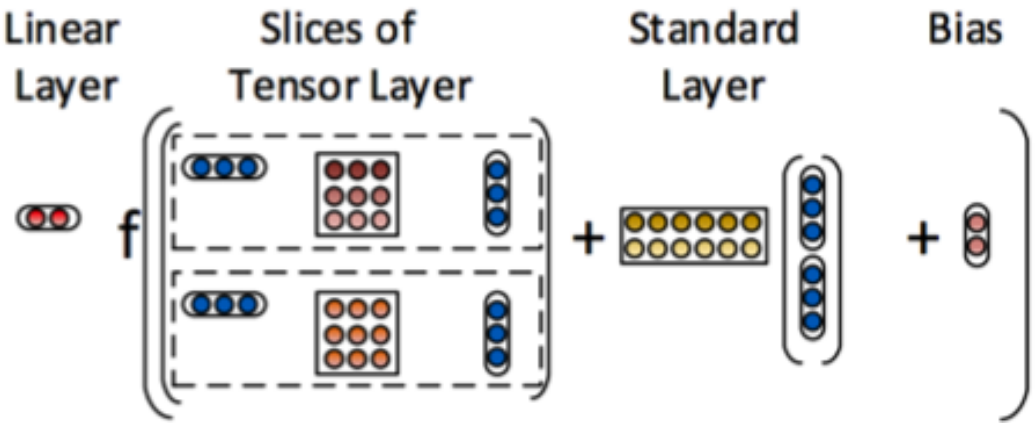
- SE also involves many parameters.
- Bad scalability w.r.t. n_r.
- Potential training problems for SE (overfitting).

Neural Tensor Networks (Socher et al., '13)

- In NTN, a relationship is represented by a tensor, 2 matrices and 2 vectors + a non-linearity (*tanh*).

$$d(sub, rel, obj) = \mathbf{u}_r^\top \tanh\left(\mathbf{h}^\top \mathcal{W}_r \mathbf{t} + \mathbf{V}_r^1 \mathbf{h} + \mathbf{V}_r^2 \mathbf{t} + \mathbf{b}_r\right)$$

- Neural Tensor layer:



- Very powerful model with high capacity for each relation.

Scalability

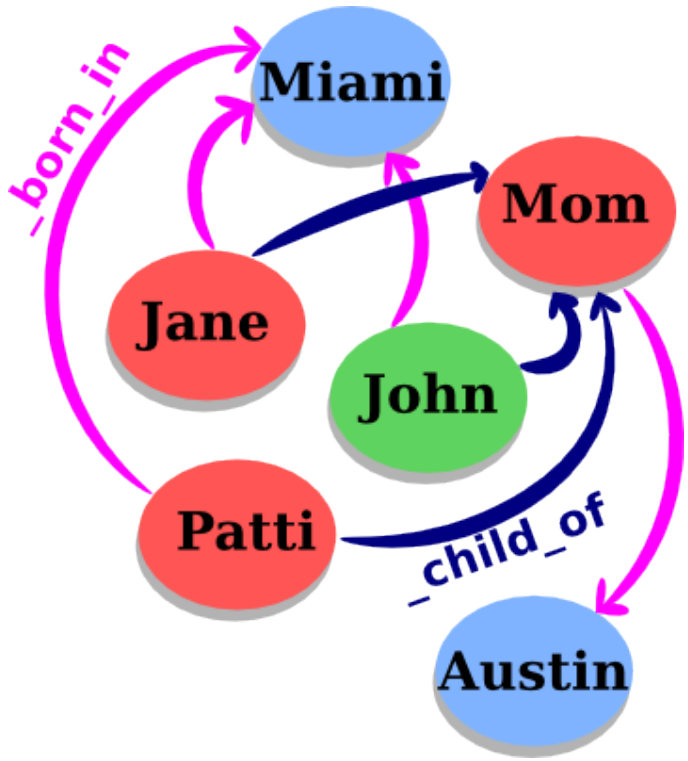
Method	**Nb of parameters**	**on Freebase15k**
RESCAL	$O(n_e d + n_r d^2)$	88M ($d = 250$)
SE	$O(n_e d + 2n_r d^2)$	8M ($d = 50$)
NTN	$O(n_e d + n_r d^3)$	165M ($d = 50$)

Freebase15k: $n_e = 15k$, $n_r = 1.3k$.

- Very high modeling capacity.
- Involves many parameters.
- Bad scalability w.r.t. n_r (overfitting if few triples).

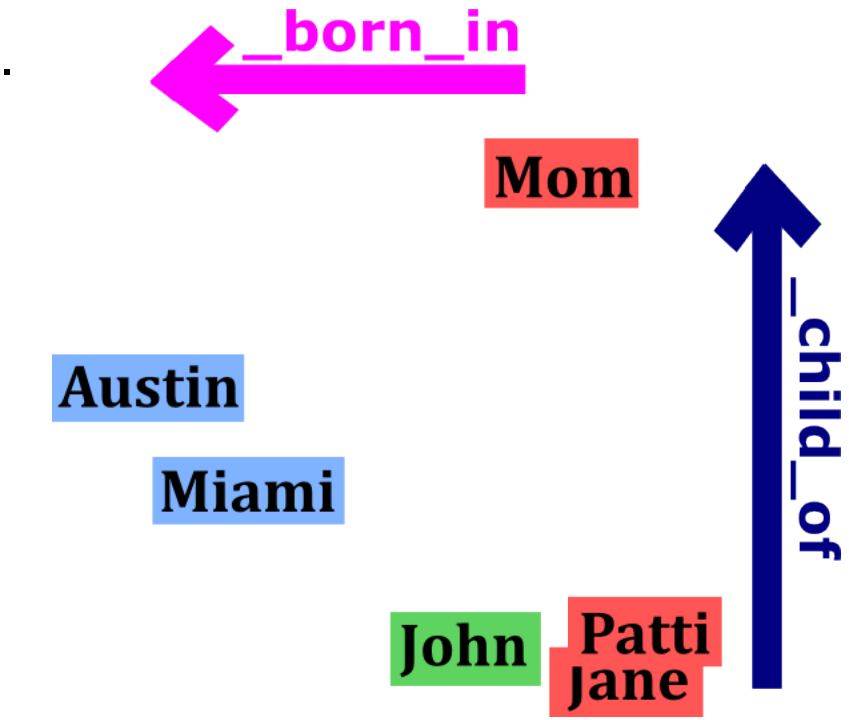
Modeling Relations as Translations (Bordes et al. '13)

Intuition: we want $\mathbf{s} + \mathbf{r} \approx \mathbf{o}$.



Modeling Relations as Translations (NIPS13)

Intuition: we would like that $\mathbf{s} + \mathbf{r} \approx \mathbf{o}$.



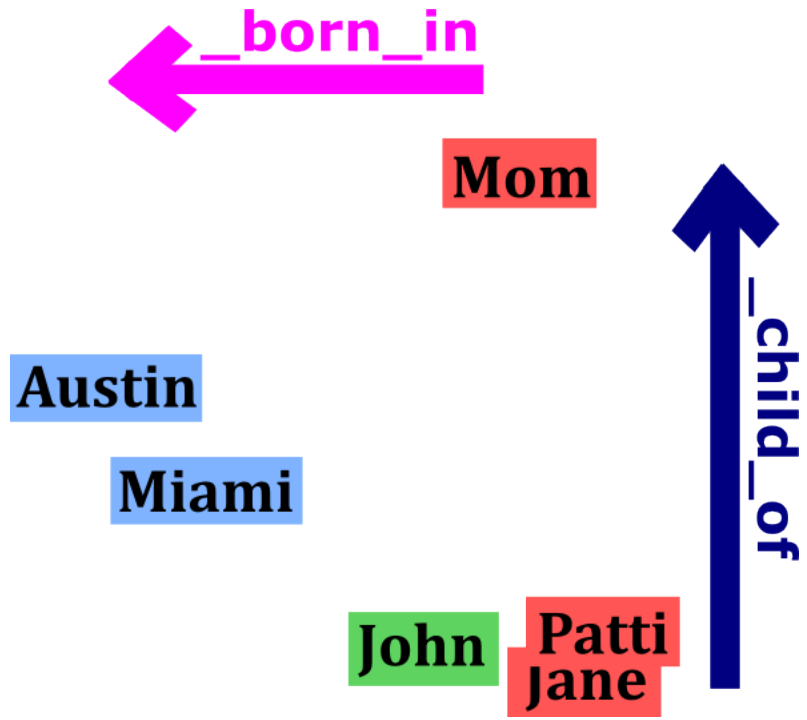
Modeling Relations as Translations (Bordes et al. '13)

Intuition: we want $\mathbf{s} + \mathbf{r} \approx \mathbf{o}$.

The similarity measure is defined as:

$$d(sub, rel, obj) = ||\mathbf{s} + \mathbf{r} - \mathbf{o}||_2^2$$

s,**r** and **o** are learned to verify that.



Learning TransE

For training, a margin ranking criterion is minimized:

$$\sum_{pos} \sum_{neg \in S'} \left[\gamma + ||\mathbf{s} + \mathbf{r} - \mathbf{o}||_2^2 - ||\mathbf{s'} + \mathbf{r} - \mathbf{o'}||_2^2 \right]_+$$

where $[x]_+$ is the positive part of x, $\gamma > 0$ is a margin, and:

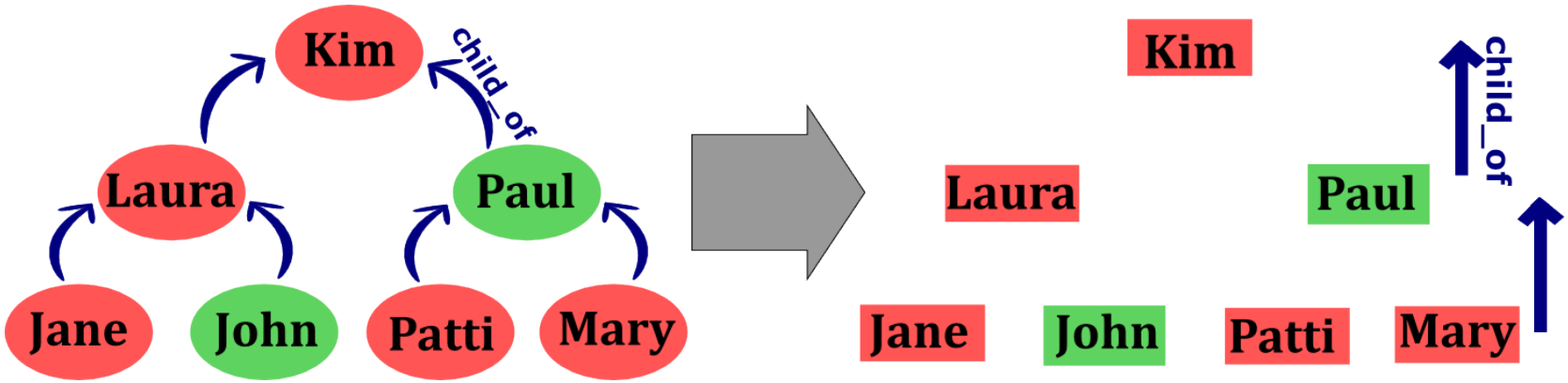
$$S' = \left\{ (\text{sub}', \text{rel}, \text{obj}) | \text{sub}' \in \mathcal{E} \right\} \cup \left\{ (\text{sub}, \text{rel}, \text{obj}') | \text{obj}' \in \mathcal{E} \right\}$$

Learning TransE

input: Training set $S = \{(\text{sub,rel,obj})\}$, margin γ, learning rate λ
initialize $\mathbf{r} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each rel
    $\mathbf{r} \leftarrow \ell / \|\ell\|$ for each ℓ
    $\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{j}}, \frac{6}{\sqrt{k}})$ for each entity ent(sub or obj)
loop
    $\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity ent
    $S_{batch} \leftarrow sample(S, b)$ //sample minibatch of size b
    $T_{batch} \leftarrow \emptyset$ //initialize set of pairs
    for (sub,rel,obj) $\in S_{batch}$ **do**
        (sub′,rel,obj′) $\leftarrow$sample(S'(sub,rel,obj)) //sample negative triplet
        $T_{batch} \leftarrow T_{batch} \cup \{((\text{sub,rel,obj}), (\text{sub}',\text{rel,obj}'))\}$
    end for
    Update embeddings w.r.t. $\sum_{T_{batch}} \nabla \left[\gamma + ||\mathbf{s} + \mathbf{r} - \mathbf{o}||_2^2 - ||\mathbf{s'} + \mathbf{r} - \mathbf{o'}||_2^2\right]_+$
end loop

Motivations of a Translation-based Model

- Natural representation for hierarchical relationships.



- Recent work on word embeddings (Mikolov et al., '13): there may exist embedding spaces in which relationships among concepts are represented by translations.

Scalability

Method	**Nb of parameters**	**on Freebase15k**
RESCAL	$O(n_e d + n_r d^2)$	88M ($d = 250$)
SE	$O(n_e d + 2n_r d^2)$	8M ($d = 50$)
NTN	$O(n_e d + n_r d^3)$	165M ($d = 50$)
TransE	$O(n_e d + n_r d)$	0.8M ($d = 50$)

Freebase15k: $n_e = 15k$, $n_r = 1.3k$.

- TransE is a special case of SE and NTN.
- TransE obtains better training errors: less overfitting.
- Much better scalability.

Chunks of Freebase

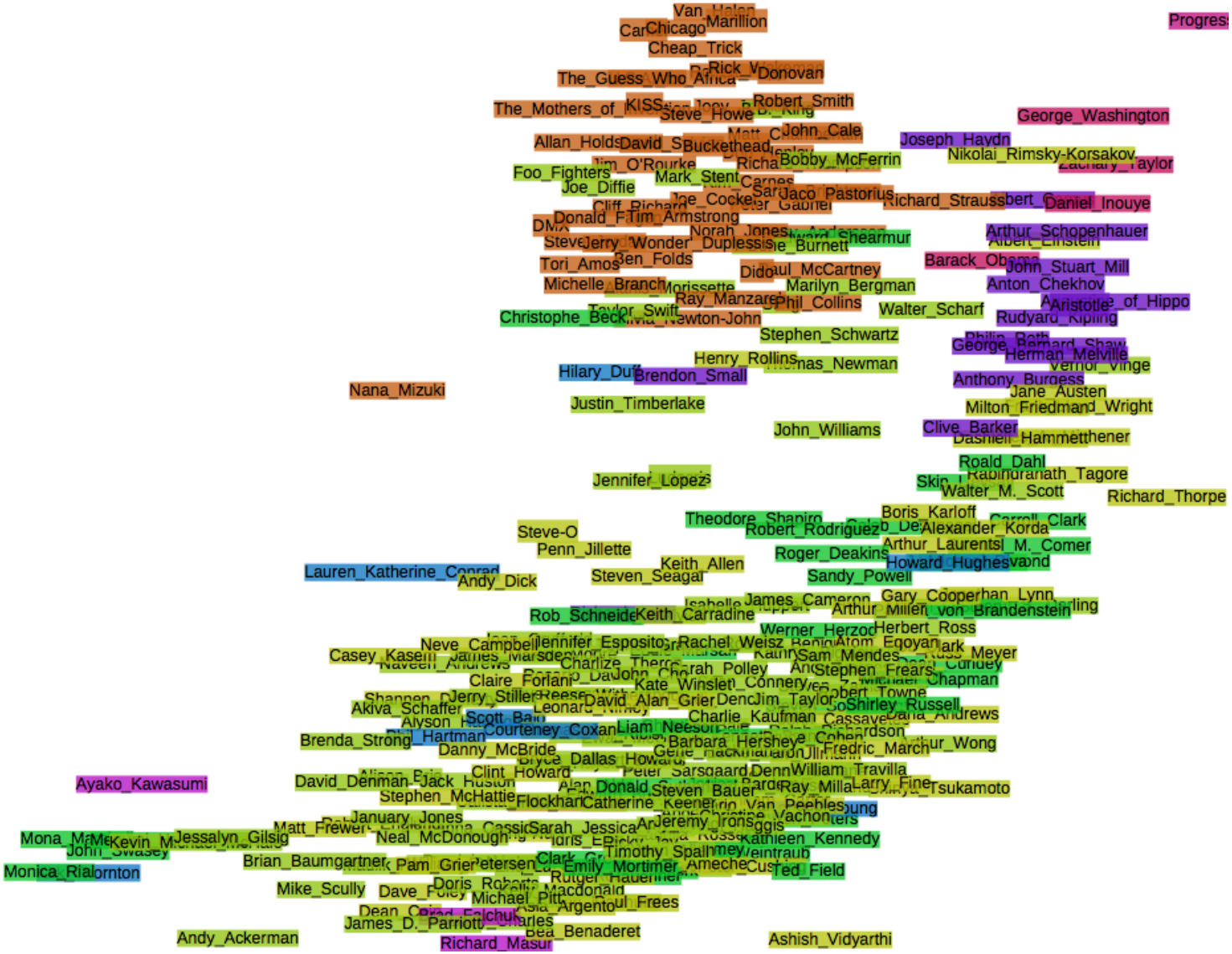
- **Data statistics**:

	Entities (n_e)	Rel. (n_r)	Train. Ex.	Valid. Ex.	Test Ex.
FB13	75,043	13	316,232	5,908	23,733
FB15k	14,951	1,345	483,142	50,000	59,071
FB1M	1×10^6	23,382	17.5×10^6	50,000	177,404

- **Training times for** TransE:
  - Embedding dimension: 50.
  - Training time:
    - on Freebase15k: ≈2h (on 1 core),
    - on Freebase1M: ≈1d (on 16 cores).

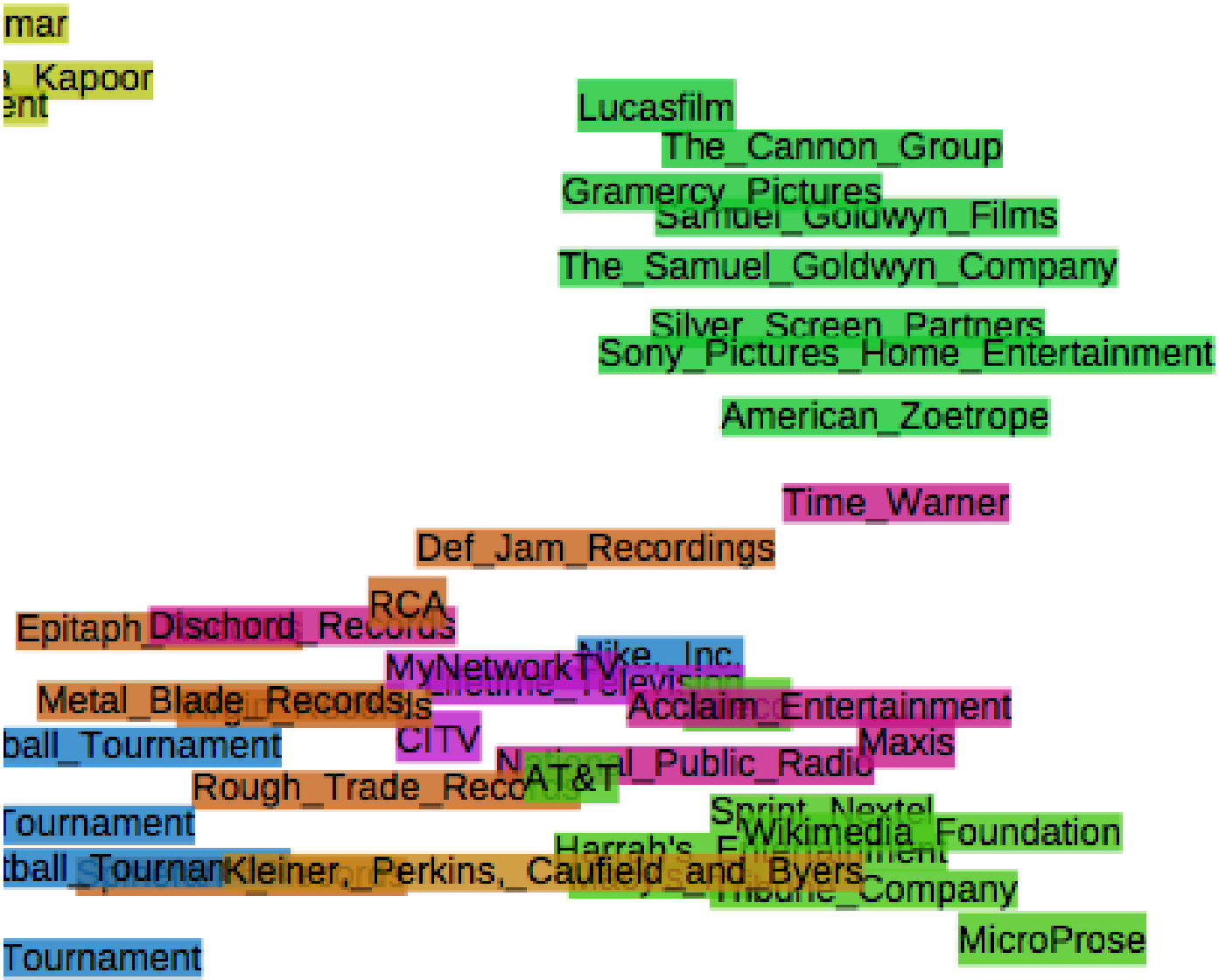
Visualization of 1,000 Entities

Visualization of 1,000 Entities - Zoom 1



Visualization of 1,000 Entities - Zoom 2

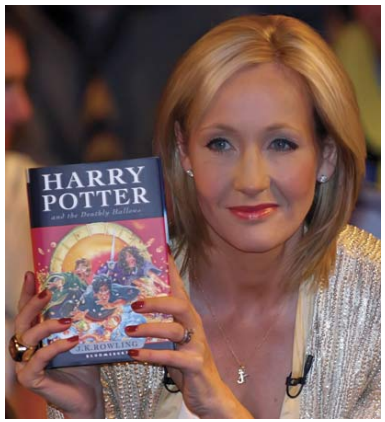
Visualization of 1,000 Entities - Zoom 3



Example

"Who influenced J.K. Rowling?"

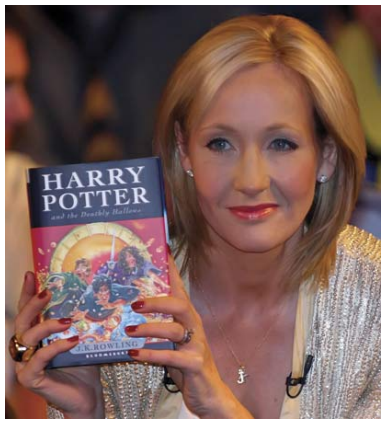
J. K. Rowling `_influenced_by` ?



Example

"Who influenced J.K. Rowling?"

J. K. Rowling `_influenced_by`



- G. K. Chesterton
- J. R. R. Tolkien
- C. S. Lewis
- Lloyd Alexander
- Terry Pratchett
- Roald Dahl
- Jorge Luis Borges
- Stephen King
- Ian Fleming

Example

"Which genre is the movie WALL-E?"

WALL-E `_has_genre` ?

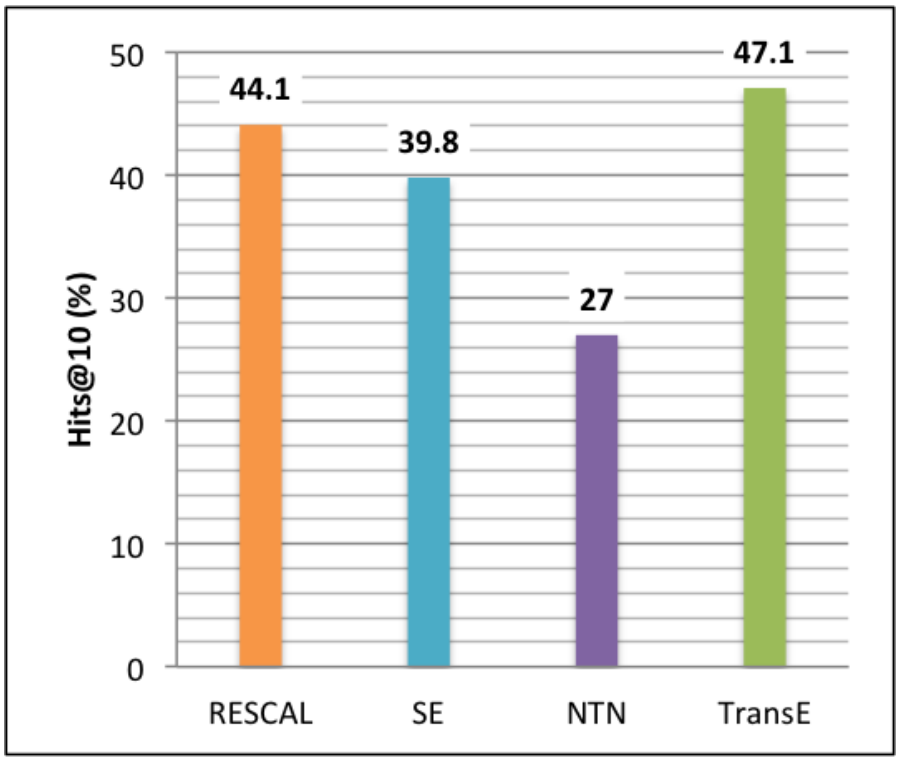
Example

"Which genre is the movie WALL-E?"

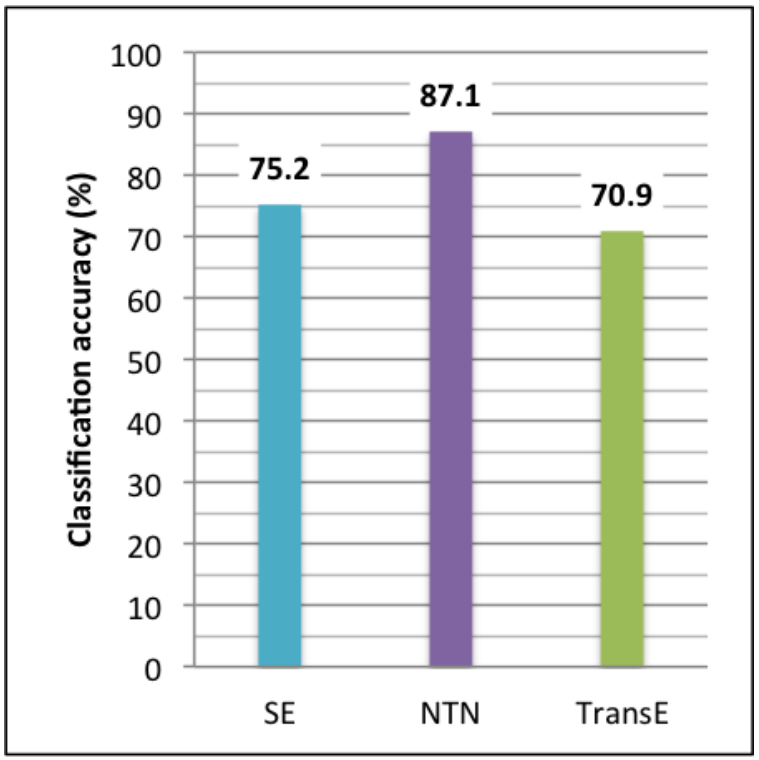
WALL-E	`_has_genre`	Animation
		Computer animation
		Comedy film
		Adventure film
		Science Fiction
		Fantasy
		Stop motion
		Satire
		Drama

Benchmarking

Ranking on FB15k



Classification on FB13



On FB1M,TransE predicts 34% in the Top-10 (SE only 17.5%).

Results extracted from (Bordes et al., '13) and (Wang et al., '14)

Refining TransE

- TATEC (García-Durán et al., '14) supplements TransE with a trigram term for encoding complex relationships:

$$d(sub, rel, obj) = \overbrace{\mathbf{s}_1^\top \mathbf{R}\mathbf{o}_1}^{\text{trigram}} + \overbrace{\mathbf{s}_2^\top \mathbf{r} + \mathbf{o}_2^\top \mathbf{r}' + \mathbf{s}_2^\top \mathbf{D}\mathbf{o}_2}^{\text{bigrams}\approx\text{TransE}},$$

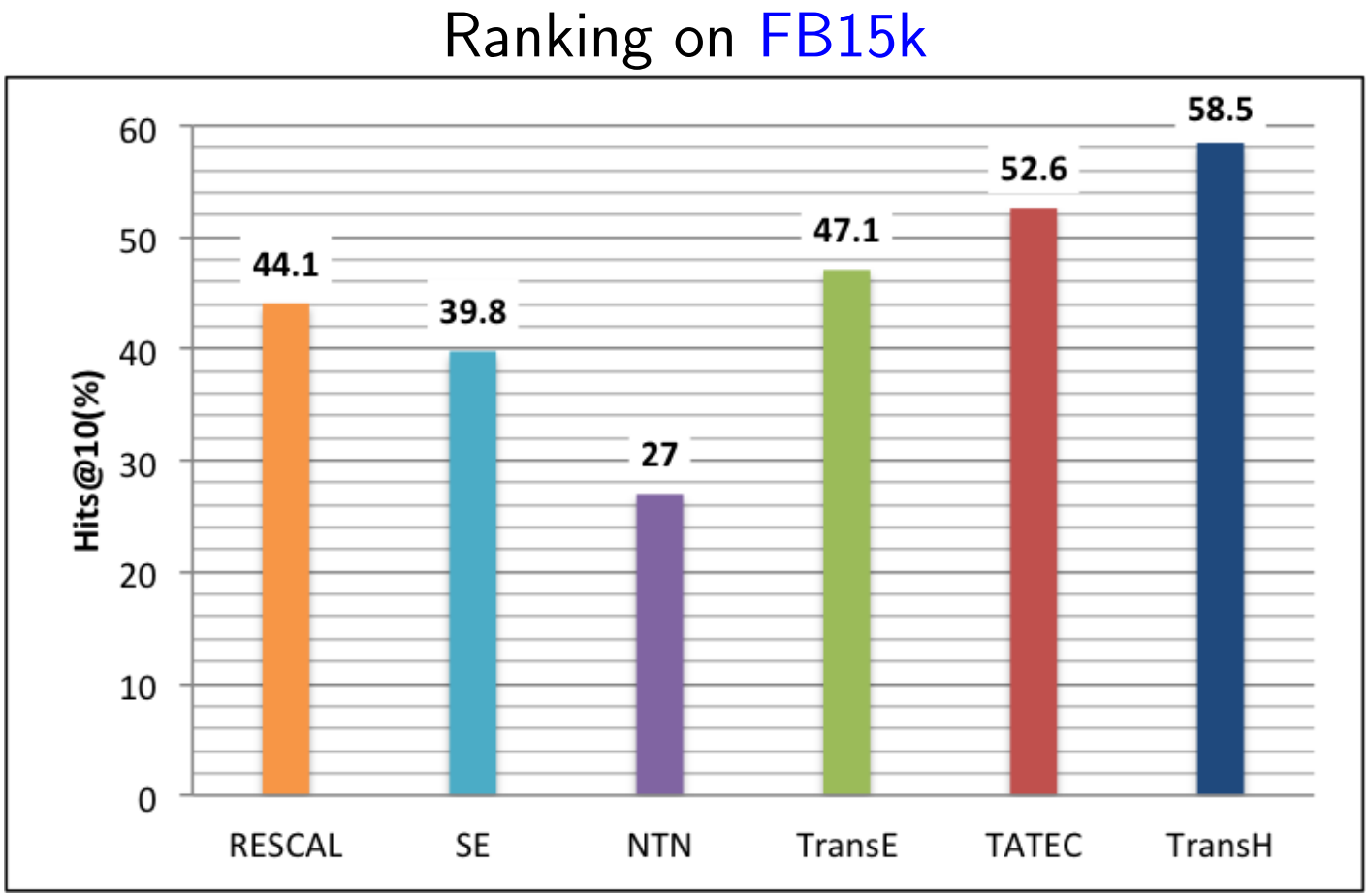
with $\mathbf{s}_1 \neq \mathbf{s}_2$ and $\mathbf{o}_1 \neq \mathbf{o}_2$.

- TransH (Wang et al., '14) adds an orthogonal projection to the translation of TransE:

$$d(sub, rel, obj) = ||(\mathbf{s} - \mathbf{r}_p^\top \mathbf{s}\mathbf{r}_p) + \mathbf{r}_t - (\mathbf{o} - \mathbf{r}_p^\top \mathbf{o}\mathbf{r}_p)||_2^2,$$

with $\mathbf{r}_p \perp \mathbf{r}_t$.

Benchmarking



Results extracted from (García-Durán et al., '14) and (Wang et al., '14)

Menu – Part 2

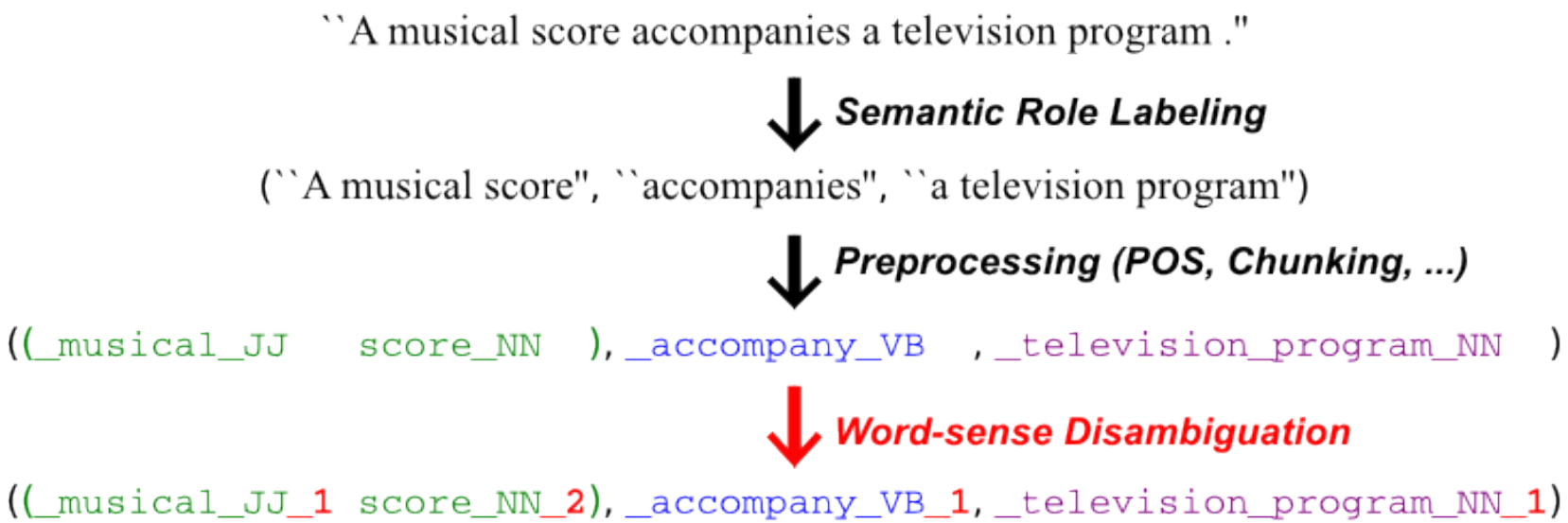
Information Extraction

- Information extraction: populate KBs with new facts using text
- Usually two steps:
  - **Entity linking**: identify mentions of entities in text
  - **Relation extraction**: extract facts about them
- Previous works include rule-based models, classifiers with features from parsers, graphical models, etc.
- Embedding models exist for both steps.

Entity Linking as WSD

Word Sense Disambiguation $\leftrightarrow$ WordNet entity linking

Towards open-text semantic parsing:



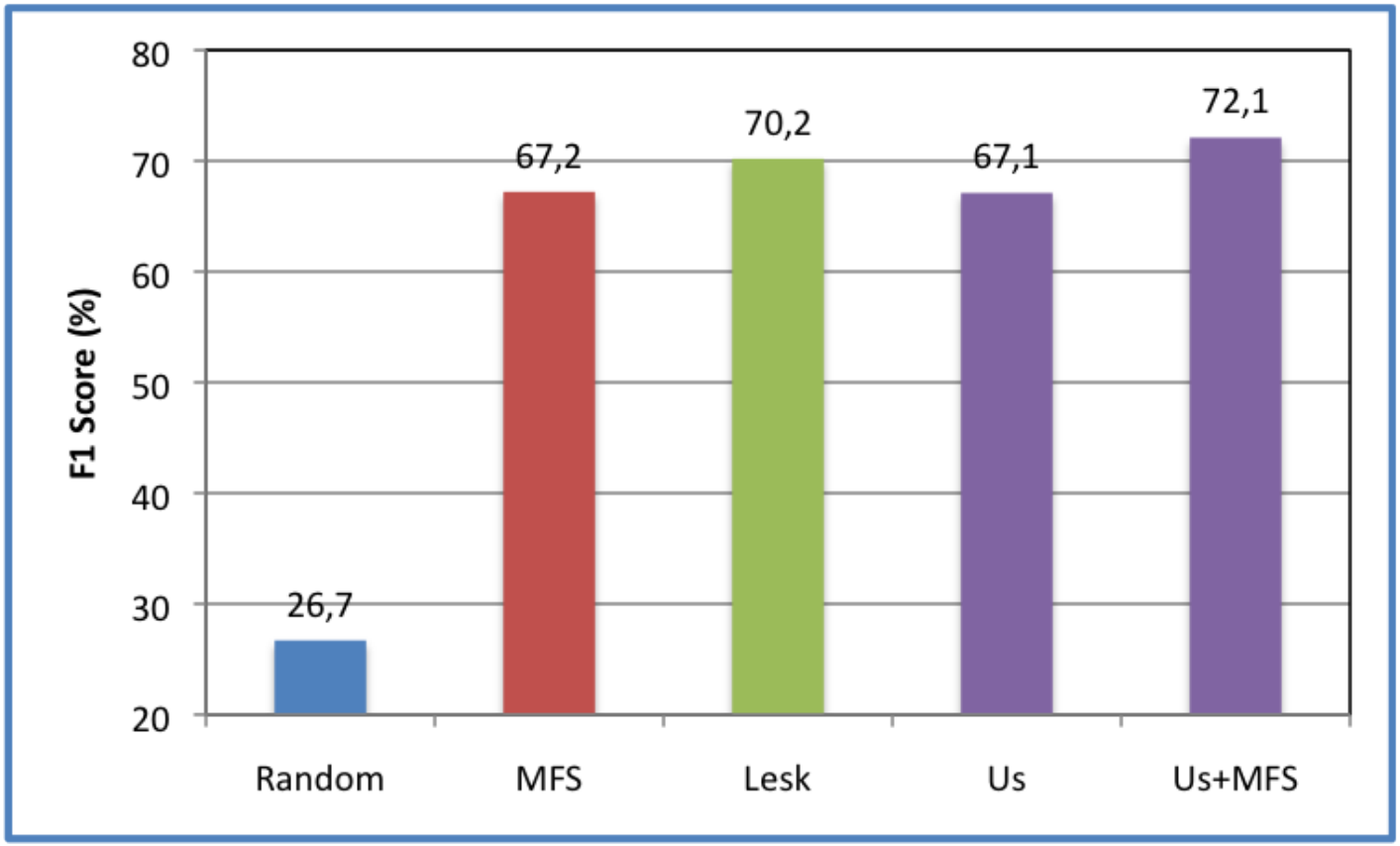
Embeddings of Text and WordNet (Bordes et al., '12)

- Text is converted into relations (*sub*,*rel*,*obj*).
- Joint learning of embeddings for all symbols: words, entities and relation types from WordNet.
- This system can label 37,141 words with 40,943 synsets.

	Train. Ex.	**Test Ex.**	**Labeled?**	**Symbol**
WordNet	146,442	5,000	No	synsets
Wikipedia	2,146,131	10,000	No	words
ConceptNet	11,332	0	Non	words
Ext. WordNet	42,957	5,000	Yes	words+synsets
Unamb. Wikip.	981,841	0	Yes	words+synsets
TOTAL	3,328,703	20,000	-	-

Benchmarking on Extended WordNet

F1-score on 5,000 test sentences to disambiguate.



Results extracted from (Bordes et al., '12)

WordNet is enriched through text

Similarities among senses beyond WordNet

"what does an army attack?"

army_NN_1 attack_VB_1 ?

WordNet is enriched through text

Similarities among senses beyond original WordNet data

"what does an army attack?"

army_NN_1	attack_VB_1	troop_NN_4
		armed_service_NN_1
		ship_NN_1
		territory_NN_1
		military_unit_NN_1

WordNet is enriched through text

Similarities among senses beyond WordNet

"Who or what earns money"

? `earn_VB_1` money_NN_1

WordNet is enriched through text

Similarities among senses beyond original WordNet data

"Who or what earns money"

person_NN_1 `earn_VB_1` money_NN_1
business_firm_NN_1
family_NN_1
payoff_NN_3
card_game_NN_1

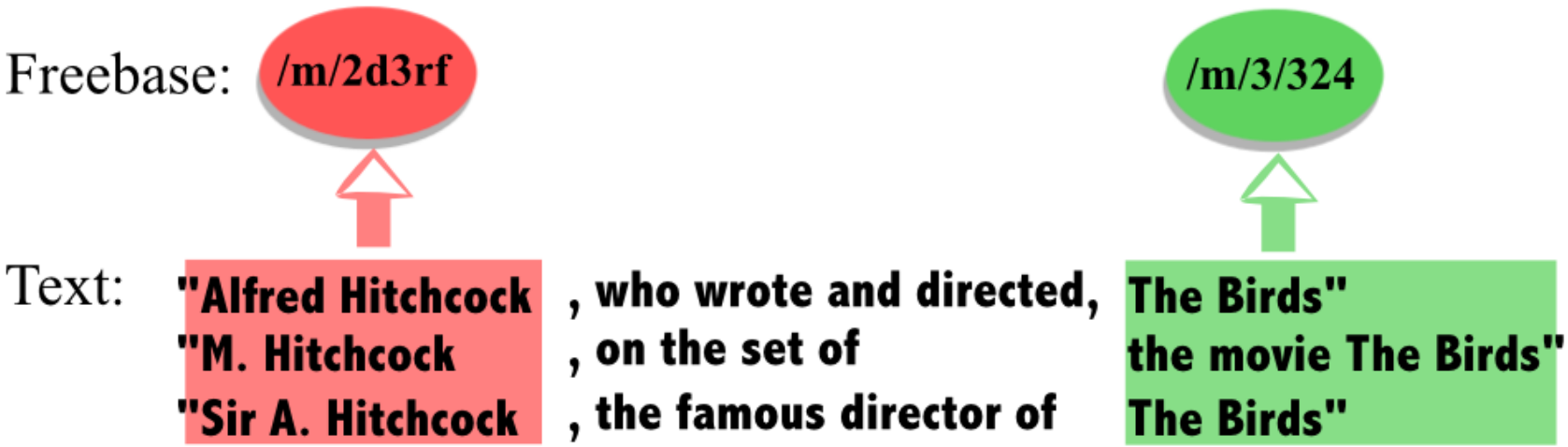
Relation Extraction

Given a bunch of sentences.

Text:	"Alfred Hitchcock	, who wrote and directed,	The Birds"
	"M. Hitchcock	, on the set of	the movie The Birds"
	"Sir A. Hitchcock	, the famous director of	The Birds"

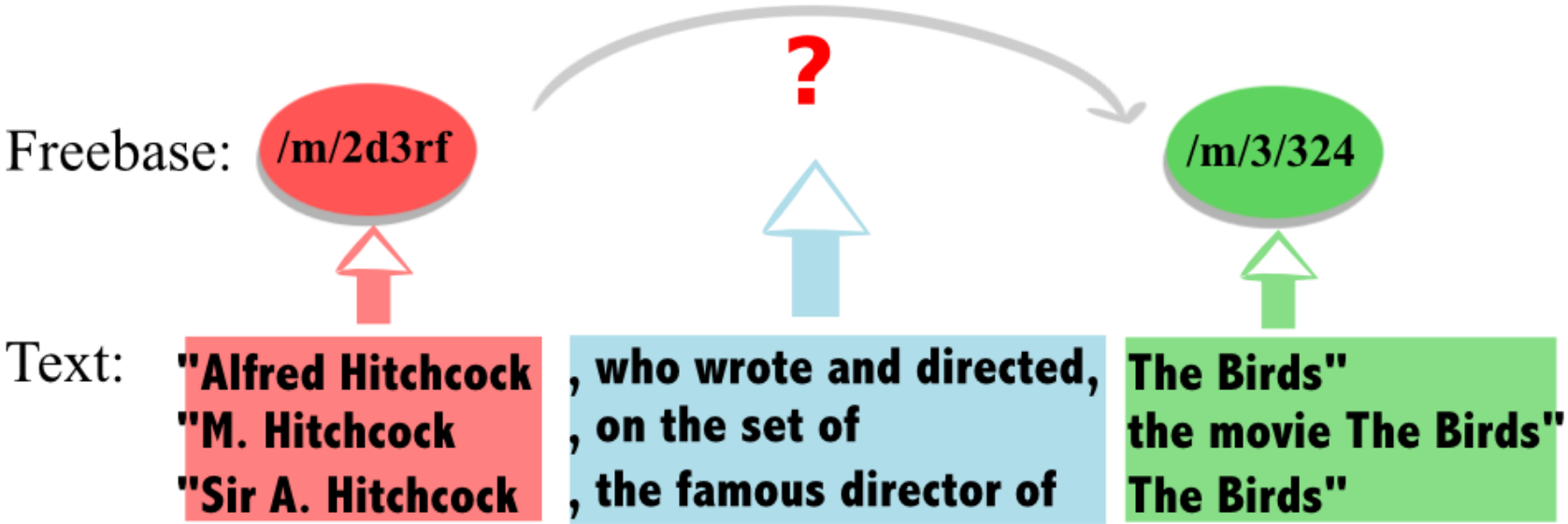
Relation Extraction

Given a bunch of sentences concerning the same pair of entities.



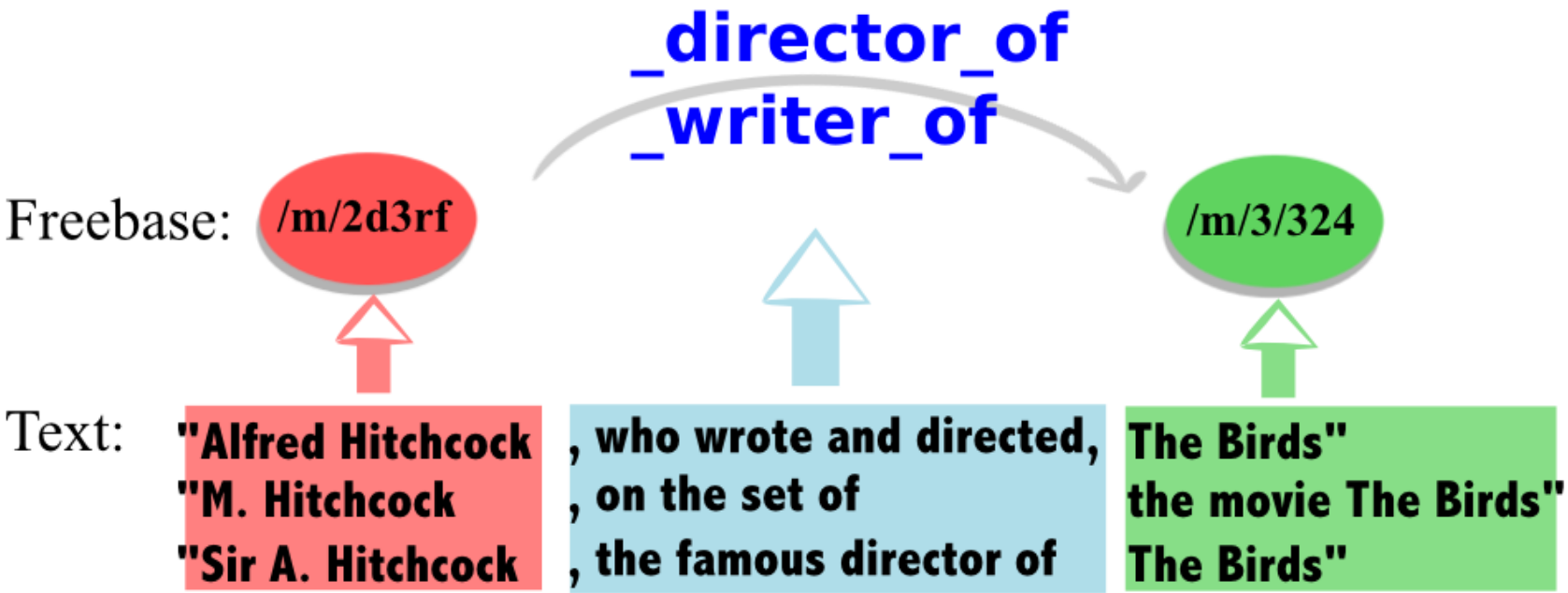
Relation Extraction

Goal: identify if there is a relation between them to add to the KB.



Relation Extraction

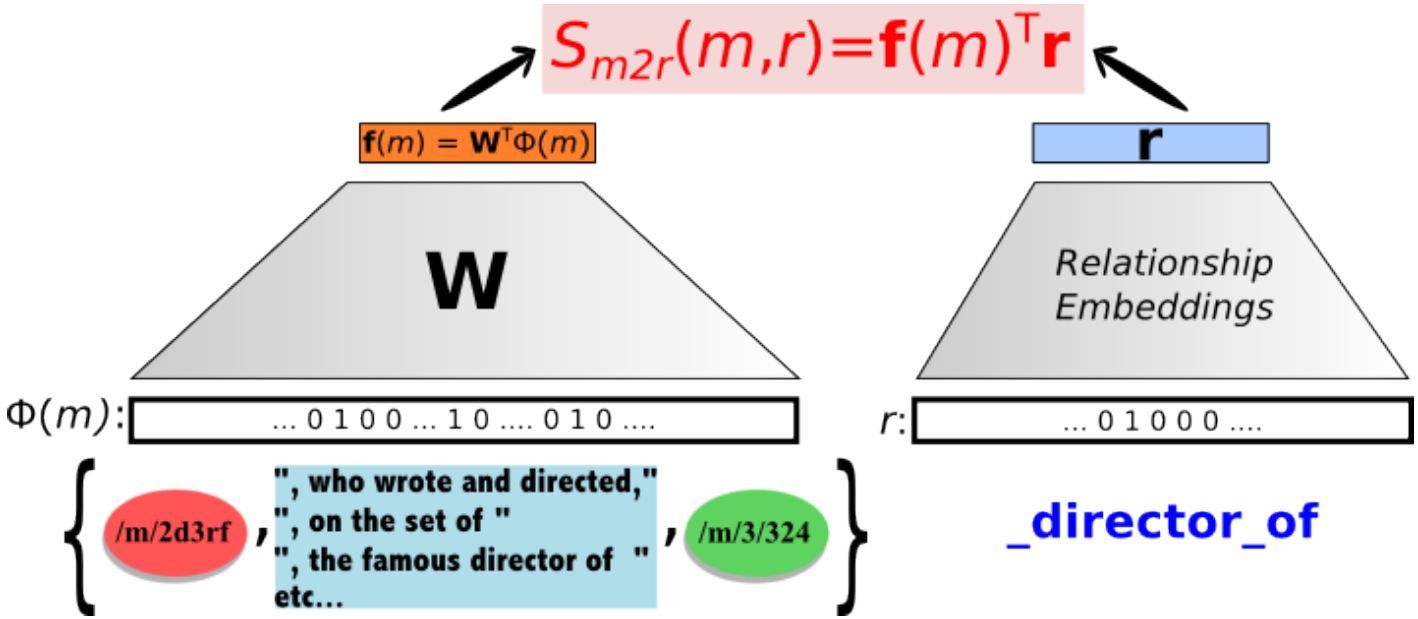
And from which type, to enrich an existing KB.



Embeddings of Text and Freebase (Weston et al., '13)

- **Standard Method:** an embedding-based classifier is trained to predict the relation type, given text mentions $\mathcal{M}$ and (sub, obj):

$$r(m, sub, obj) = \arg\max_{rel'} \sum_{m \in \mathcal{M}} S_{m2r}(m, rel')$$



Classifier based on WSABIE (Weston et al., '11).

Embeddings of Text and Freebase (Weston et al., '13)

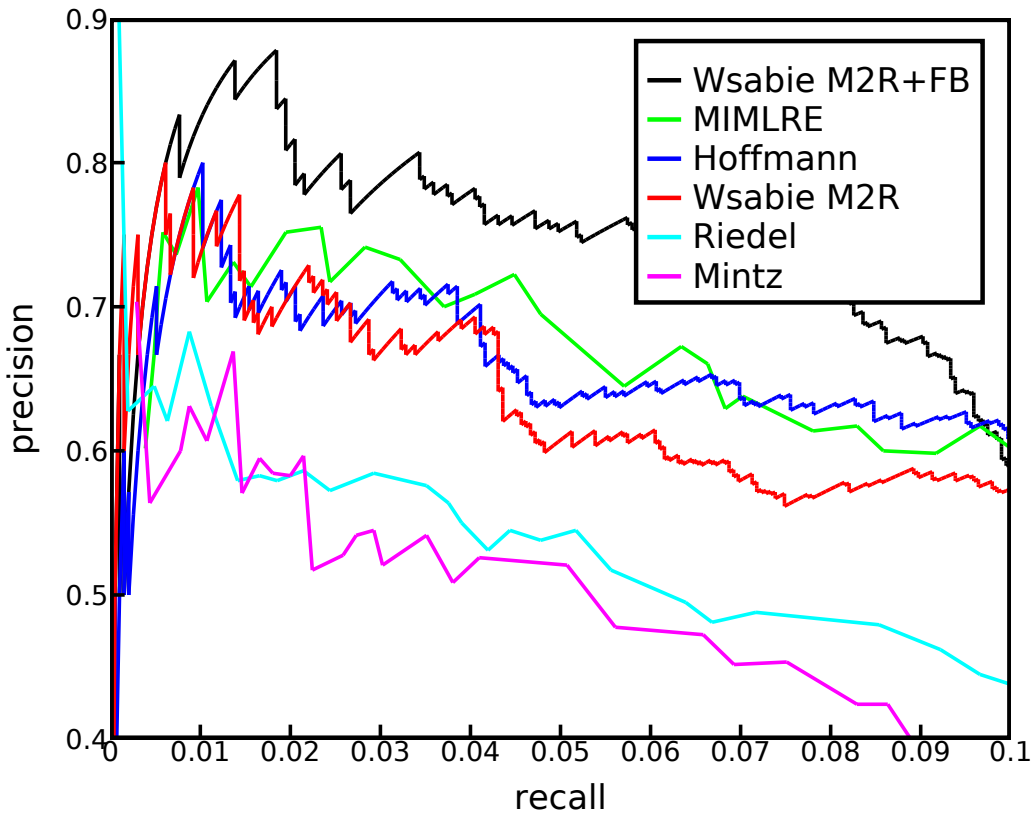
- **Idea:** improve extraction by using both text + available knowledge (= current KB).
- A model of the KB is used in a re-ranking setting to force extracted relations to agree with it:

$$r'(m, sub, obj) = \arg\max_{rel'} \Big(\sum_{m \in \mathcal{M}} S_{m2r}(m, rel') - d_{KB}(sub, rel', obj) \Big)$$

with $d_{KB}(sub, rel', obj) = ||\mathbf{s} + \mathbf{r}' - \mathbf{o}||_2^2$ (trained separately)

Benchmarking on NYT+Freebase

Exp. on NY Times papers linked with Freebase (Riedel et al., '10)

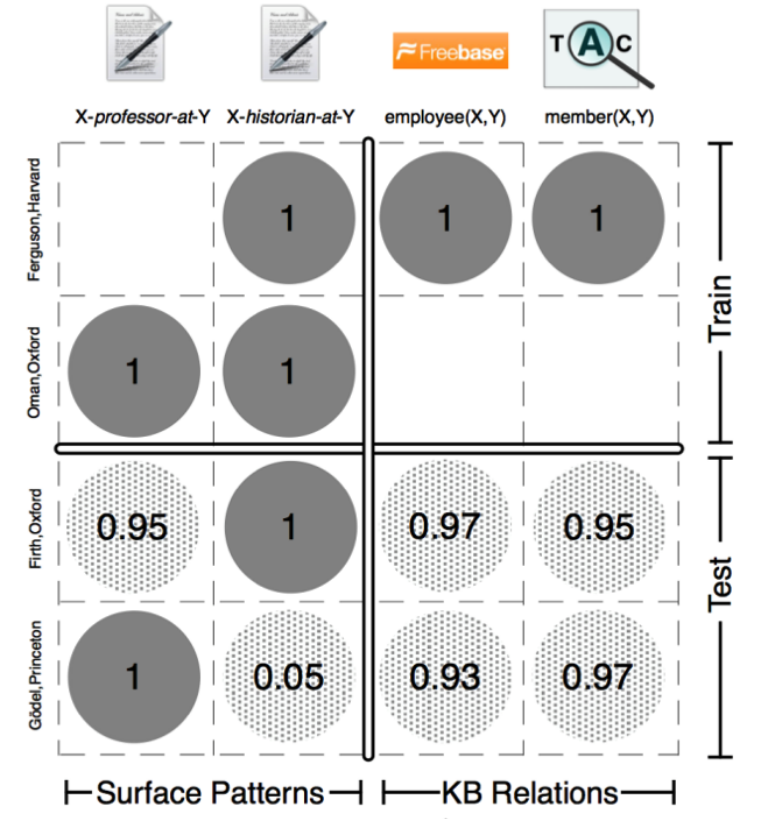


Precision/recall curve for predicting relations

Results extracted from (Weston et al., '13)

Universal Schemas (Riedel et al., '13)

- Join in a single learning problem:
  - relation extraction
  - link prediction
- The same model score triples:
  - made of text mentions
  - from a KB



Universal Schemas (Riedel et al., '13)

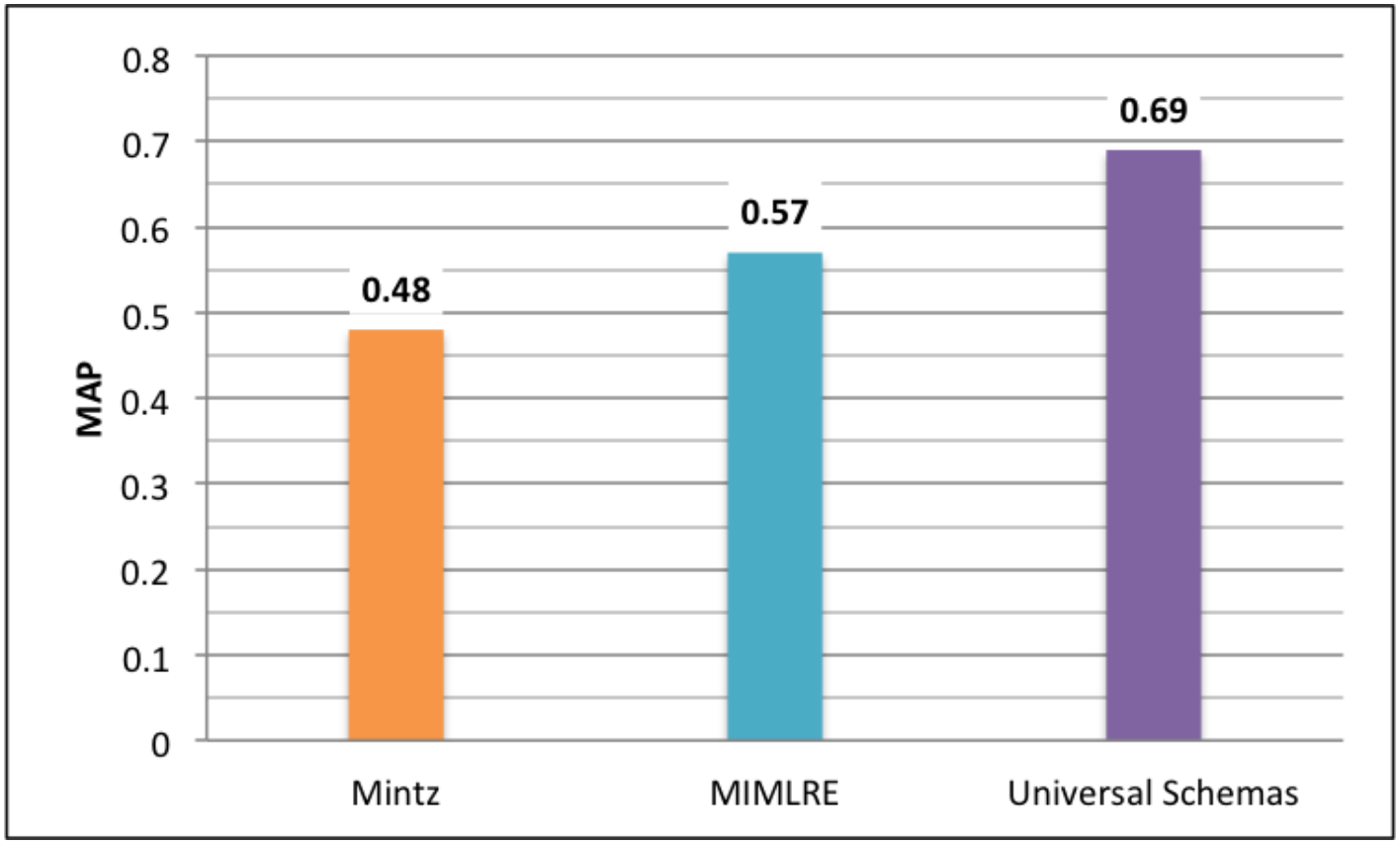
- Relation prediction using the score:

$$r'(m, sub, obj) = \arg\max_{rel'} \Big(\sum_{m \in \mathcal{M}} S_{m2r}(m, rel') + S_{KB}(sub, rel', obj) + S_{neighbors}(sub, rel', obj) \Big)$$

- All scores are defined using embeddings:
  - $S_{m2r}(m, rel') = \mathbf{f}(m)^\top \mathbf{r}'$
  - $S_{kb}(sub, rel', obj) = \mathbf{s}^\top \mathbf{r}'_s + \mathbf{o}^\top \mathbf{r}'_o$
  - $S_{neighbors}(sub, rel', obj) = \sum_{\substack{(sub, rel'', obj) \\ rel'' \neq rel'}} w^{rel'}_{rel''}$
- Training by ranking observed facts versus other and updating using SGD.

Benchmarking on NYT+Freebase

Exp. on NY Times papers linked with Freebase (Riedel et al., '10)



Mean Averaged Precision for predicting relations

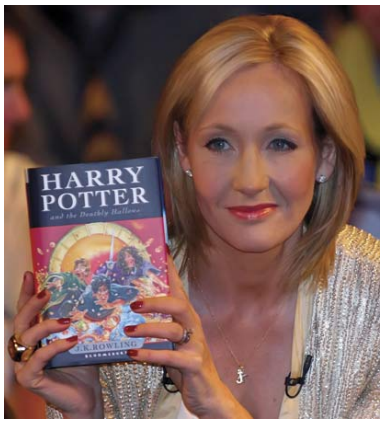
Results extracted from (Riedel et al., '13)

Menu – Part 2

Link Prediction as Q&A

"Who influenced J.K. Rowling?"

J. K. Rowling `_influenced_by`



- G. K. Chesterton
- J. R. R. Tolkien
- C. S. Lewis
- Lloyd Alexander
- Terry Pratchett
- Roald Dahl
- Jorge Luis Borges

Can we go beyond such rigid structure?

Open-domain Question Answering

- **Open-domain Q&A**: answer question on any topic
  $\longrightarrow$ query a KB with natural language

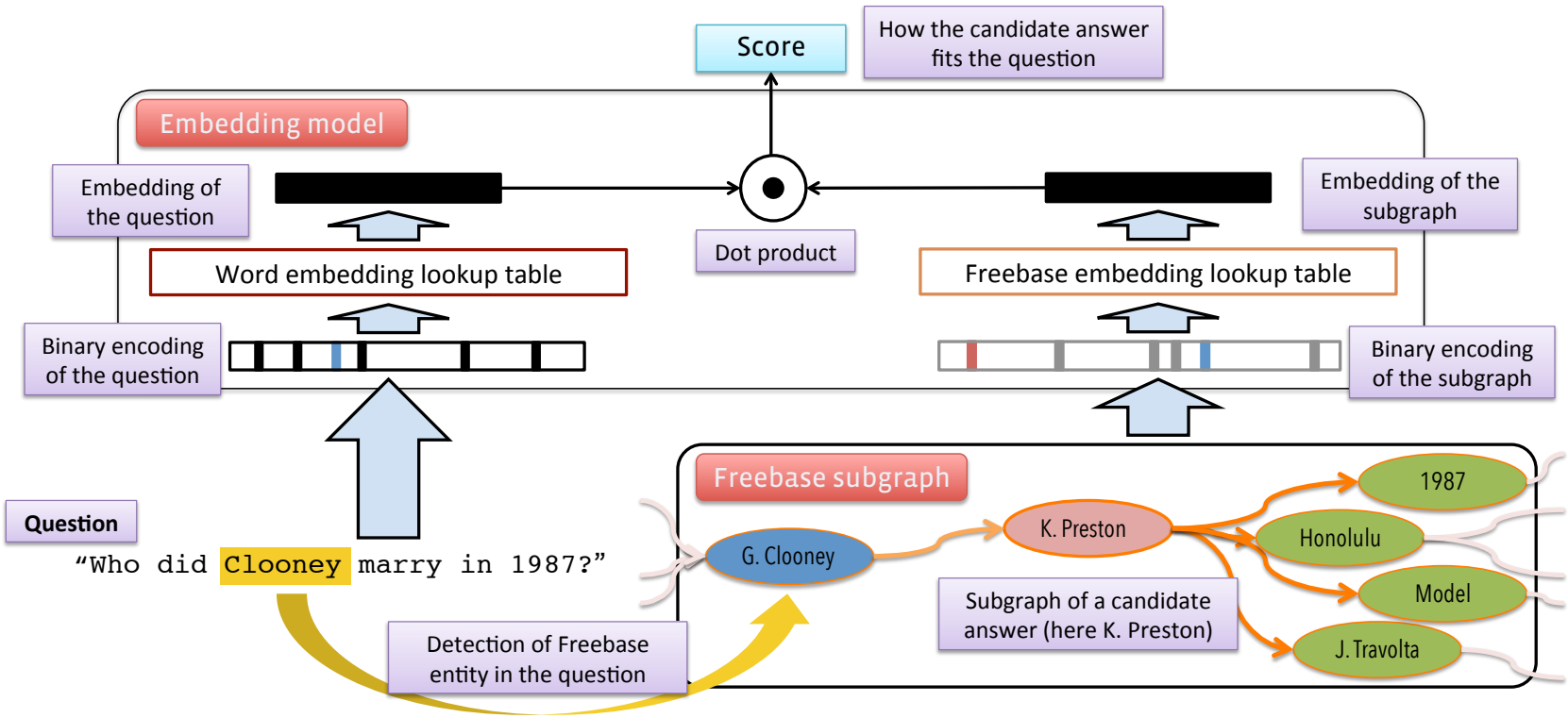
Examples

"What is `cher`'s son's name ?"	`elijah_blue_allman`
"What are `dollars` called in `spain` ?"	`peseta`
"What is `henry_clay` known for ?"	`lawyer`
"Who did `georges_clooney` marry in `1987` ?"	`kelly_preston`

- Recent effort with semantic parsing (Kwiatkowski et al. '13) (Berant et al. '13, '14) (Fader et al., '13, '14) (Reddy et al., '14)
- Models with embeddings as well (Bordes et al., '14)

Subgraph Embeddings (Bordes et al., '14)

- Model learns embeddings of questions and (candidate) answers
- Answers are represented by entity and its neighboring subgraph



Training data

- Freebase is automatically converted into Q&A pairs
- Closer to expected language structure than triples

Examples of Freebase data

(sikkim, location.in_state.judicial_capital, gangtok)
what is the judicial capital of the in state sikkim ? – gangtok

(brighouse, location.location.people_born_here, edward_barber)
who is born in the location brighouse ? – edward_barber

(sepsis, medicine.disease.symptoms, skin_discoloration)
what are the symptoms of the disease sepsis ? – skin_discoloration

Training data

- All Freebase questions have rigid and similar structures
- Supplemented by pairs from clusters of paraphrase questions
- Multitask training: similar questions $\leftrightarrow$ similar embeddings

Examples of paraphrase clusters

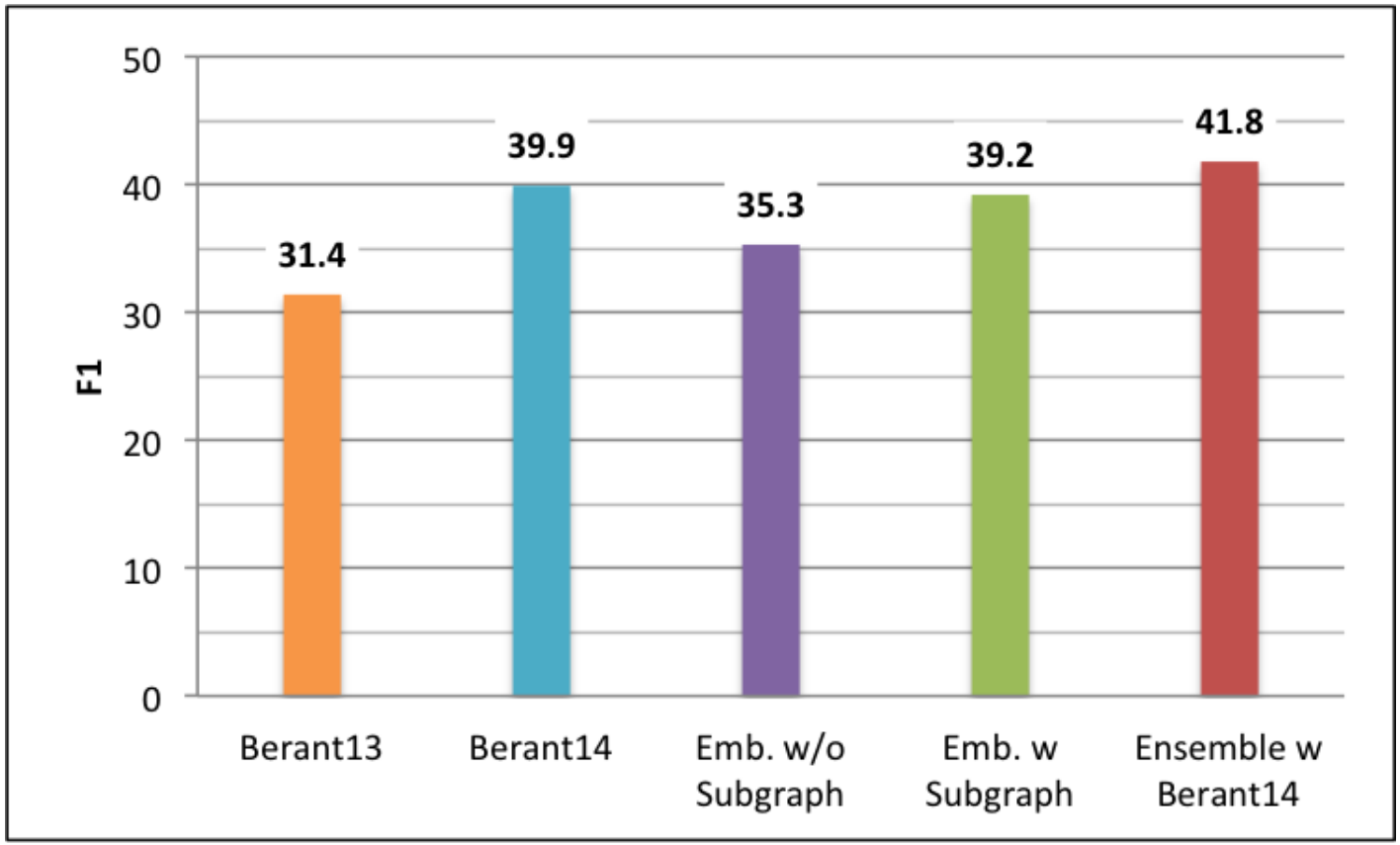
what are two reason to get a 404 ?
what is error 404 ?
how do you correct error 404 ?

what is the term for a teacher of islamic law ?
what is the name of the religious book islam use ?
who is chief of islamic religious authority ?

what country is bueno aire in ?
what countrie is buenos aires in ?
what country is bueno are in ?

Benchmarking on WebQuestions

Experiments on WebQuestions (Berant et al., '13)



F1-score for answering test questions

Results extracted from (Berant et al., '14) and (Bordes et al., '14)

Menu – Part 2

Advantages

- Efficient features for many tasks in practice
- Training with SGD scales & parallelizable (Niu et al., '11)
- Flexible to various tasks: multi-task learning of embeddings
- Supervised or unsupervised training
- Allow to use extra-knowledge in other applications

Issues

- Must train all embeddings together (no parallel 1-vs-rest)
- Low-dimensional vector $\longrightarrow$ compression, blurring
- Sequential models suffer from long-term memory
- Embeddings need quite some updates to be good – not 1-shot
- Negative example sampling can be unefficient

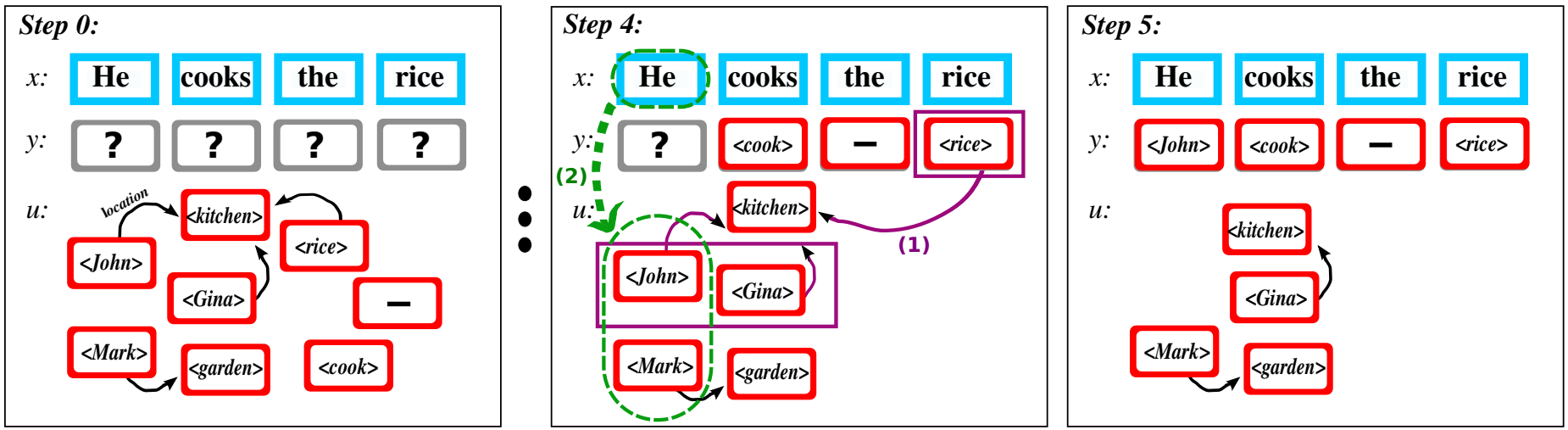
Menu – Part 2

Fix current limitations

- Compression: improve the memory capacity of embeddings and allows for one-shot learning of new symbols
- Long-term memory: encode longer dependencies in sequential models like RNNs
- Training: faster and better sampling of examples
- Beyond linear: most supervised labeling problems are well tackled by simple linear models. Non-linearity should help more.

Explore new directions

- Compositionality (Baroni et al. '10) (Grefenstette, 13)
- Multimodality (Bruni et al., 12) (Kiros et al., '14)
- Grounding language into actions (Bordes et al., 10)



At EMNLP

Modeling Interestingness with Deep Neural Networks
Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He and Li Deng

Translation Modeling with Bidirectional Recurrent Neural Networks
Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker and Hermann Ney

Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics
Douwe Kiela and Léon Bottou

Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean
Felix Hill and Anna Korhonen

Incorporating Vector Space Similarity in Random Walk Inference over Knowledge Bases
Matt Gardner, Partha Talukdar, Jayant Krishnamurthy and Tom Mitchell

Composition of Word Representations Improves Semantic Role Labelling
Michael Roth and Kristian Woodsend

At EMNLP

A Neural Network for Factoid Question Answering over Paragraphs
Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher and Hal Daumé III

Joint Relational Embeddings for Knowledge-based Question Answering
Min-Chul Yang, Nan Duan, Ming Zhou and Hae-Chang Rim

Evaluating Neural Word Representations in Tensor-Based Compositional Settings
Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh and Matthew Purver

Opinion Mining with Deep Recurrent Neural Networks
Ozan Irsoy and Claire Cardie

The Inside-Outside Recursive Neural Network model for Dependency Parsing
Phong Le and Willem Zuidema

A Fast and Accurate Dependency Parser using Neural Networks
Danqi Chen and Christopher Manning

At EMNLP

Reducing Dimensions of Tensors in Type-Driven Distributional Semantics
Tamara Polajnar, Luana Fagarasan and Stephen Clark

Word Semantic Representations using Bayesian Probabilistic Tensor Factorization
Jingwei Zhang, Jeremy Salwen, Michael Glass and Alfio Gliozzo

Glove: Global Vectors for Word Representation
Jeffrey Pennington, Richard Socher and Christopher Manning

Jointly Learning Word Representations and Composition Functions Using Predicate-Argument Structures
Kazuma Hashimoto, Pontus Stenetorp, Makoto Miwa and Yoshimasa Tsuruoka

Typed Tensor Decomposition of Knowledge Bases for Relation Extraction
Kai-Wei Chang, Wen-tau Yih, Bishan Yang and Christopher Meek

Knowledge Graph and Text Jointly Embedding
Zhen Wang, Jianwen Zhang, Jianlin Feng and Zheng Chen

At EMNLP

Question Answering with Subgraph Embeddings
Antoine Bordes, Sumit Chopra and Jason Weston

Word Translation Prediction for Morphologically Rich Languages with Bilingual Neural Networks
Ke M. Tran, Arianna Bisazza and Christof Monz

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation
Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio

Convolutional Neural Networks for Sentence Classification
Yoon Kim

#TagSpace: Semantic Embeddings from Hashtags
Jason Weston, Sumit Chopra and Keith Adams

Focus on (Kiros et al., 14)

Multimodal Neural Language Models

Ryan Kiros, Ruslan Salakhutdinov, Richard Zemel. ICML, 2014.

Menu – Part 2

Code

- Torch: `www.torch.ch`
- SENNA: `ronan.collobert.com/senna`
- RNNLM: `www.fit.vutbr.cz/~imikolov/rnnlm`
- Word2vec: `code.google.com/p/word2vec`
- Recursive NN: `nlp.stanford.edu/sentiment`
- SME (multi-relational data): `github.com/glorotxa/sme`

References

Semantic Parsing on Freebase from Question-Answer Pairs
J. Berant, A. Chou, R. Frostig & P. Liang. *EMNLP, 2013*

Semantic Parsing via Paraphrasing
J. Berant & P. Liang. *ACL, 2013*

Learning Structured Embeddings of Knowledge Bases
A. Bordes, J. Weston, R. Collobert & Y. Bengio. *AAAI, 2011*

Joint Learning of Words and Meaning Rep. for Open-Text Semantic Parsing
A. Bordes, X. Glorot, J. Weston & Y. Bengio. *AISTATS, 2012*

A Semantic Matching Energy Function for Learning with Multi-relational Data
A. Bordes, X. Glorot, J. Weston & Y. Bengio. *MLJ, 2013*

Translating Embeddings for Modeling Multi-relational Data
A. Bordes, N. Usunier, A. García-Durán, J. Weston & O. Yakhnenko. *NIPS, 2013*

References

Question Answering with Subgraph Embeddings
A. Bordes, S. Chopra & J. Weston. *EMNLP, 2014*

Clustering with Multi-Layer Graphs: A Spectral Perspective
X. Dong, P. Frossard, P. Vandergheynst & N. Nefedov. *IEEE TSP, 2013*

Paraphrase-Driven Learning for Open Question Answering
A. Fader, L. Zettlemoyer & O. Etzioni. *ACL, 2013*

Open Question Answering Over Curated and Extracted Knowledge Bases
A. Fader, L. Zettlemoyer & O. Etzioni. *KDD, 2014*

Learning Probabilistic Relational Models
N. Friedman, L. Getoor, D. Koller & A. Pfeffer. *IJCAI, 1999*

Effective Blending of Two and Three-way Interactions for Modeling Multi-relational Data
A. García-Durán, A. Bordes & N. Usunier. *ECML-PKDD, 2014*

References

Models for the Analysis of Asymmetrical Relationships among n Objects or Stimuli
R. Harshman. *Joint Symposium of Psych. and Mathematical Societies, 1978*.

PARAFAC: Parallel factor analysis
R. Harshman & M. Lundy. *Comp. Statistics and Data Analysis, 1994*

A Latent Factor Model for Highly Multi-relational Data
R. Jenatton, N. Le Roux, A. Bordes & G. Obozinski. *NIPS, 2012*.

Learning Systems of Concepts with an Infinite Relational Model
C. Kemp, J. Tenenbaum, T. Griffiths, T. Yamada & N. Ueda. *AAAI, 2006.*

Statistical Predicate Invention
S. Kok, P. Domingos. *ICML, 2007*

Scaling Semantic Parsers with On-the-fly Ontology Matching
T. Kwiatkowski, E. Choi, Y. Artzi & L. Zettlemoyer. *EMNLP, 2013*

References

Random Walk Inference and Learning in A Large Scale Knowledge Base
N. Lao, T. Mitchell & W. Cohen. *EMNLP, 2011*.

Distributed Representations of Words and Phrases and their Compositionality
T. Mikolov, I. Sutskever, K. Chen, G. Corrado & J. Dean. *NIPS, 2013*.

A Three-Way Model for Collective Learning on Multi-Relational Data
M. Nickel, V. Tresp & H.-P. Kriegel. *ICML, 2011*.

Large-scale Semantic Parsing without Question-Answer Pairs
S. Reddy, M. Lapata & M. Steedman. *TACL, 2014.*

Modeling Relations and Their Mentions without Labeled Text
S. Riedel, L. Yao and A. McCallum. *ECML-PKDD, 2010*

Relation Extraction with Matrix Factorization and Universal Schemas
S. Riedel, L. Yao, B. Marlin and A. McCallum. *HLT-NAACl, 2013*

References

Reasoning With Neural Tensor Networks for Knowledge Base Completion
R. Socher, D. Chen, C. Manning & A. Ng *NIPS, 2013*.

Modelling Relational Data using Bayesian Clustered Tensor Factorization
I. Sutskever, R. Salakhutdinov & J. Tenenbaum. *NIPS, 2009*.

Discriminative Probabilistic Models for Relational Data
B. Taskar, P. Abbeel & D. Koller. *UAI, 2002*.

Knowledge Graph Embedding by Translating on Hyperplanes
Z. Wang, J. Zhang, J. Feng & Z. Chen. *AAAI, 2014.*

Wsabie: Scaling Up To Large Vocabulary Image Annotation
J. Weston, S. Bengio & N. Usunier. *IJCAI, 2011*

Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction.
J. Weston, A. Bordes, O. Yakhnenko & N. Usunier. *EMNLP, 2013*